МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

«Нижегородский государственный университет им. Н.И. Лобачевского»

Институт информационных технологий, математики и механики

Кафедра теории вероятностей и анализа данных

Направление подготовки: «01.03.02 Прикладная математика и информатика»

ОТЧЁТ

по производственной практике

на тему:

«Оценка стоимости квартир с использованием математической статистики»

Выполнил:
студент группы 3821Б1ПМоп3
Чезганов Иван Алексеевич

Научный руководитель:
доцент кафедры ТВиАД
к.ф.-м.н. Бородина Татьяна Сергеевна

Руководитель от профильной организации:
аналитик


Акобян Арман Амаякович

Нижний Новгород
2024

Содержание

Введение

Текущая практика была посвящена исследованию рынка квартир с использованием методов математической статистики и машинного обучения. Основная цель заключалась в создании модели, способной точно оценить стоимость квартир на основе различных параметров и сравнении результатов с полученными для помещений свободного назначения.

Переход от анализа производственной и складской недвижимости к исследованию рынка квартир предоставляет уникальные возможности для расширения навыков в сфере анализа данных. Обработка данных о квартирах требует учета множества факторов, таких как расположение, площадь, состояние объекта и другие характеристики, что делает задачу более комплексной и многогранной.

В данной практике особое внимание уделялось изучению и обработке данных, предоставленных различными источниками. Анализ включал в себя выявление ключевых факторов, влияющих на стоимость квартир, очистку данных от дубликатов и пропусков, а также подготовку данных для построения моделей машинного обучения.

Используемый датасет содержал информацию о квартирах в городе Новосибирск, включая такие параметры, как количество комнат, площадь, этажность, год постройки, расстояние до центра города и другие важные характеристики. Этот подход позволил глубже погрузиться в специфику рынка жилой недвижимости и применить полученные знания на практике.

Основные задачи, поставленные в ходе данной работы, включали в себя:

- Проведение первичного анализа данных, выявление и устранение аномалий и выбросов.
- Выделение ключевых параметров, влияющих на стоимость квартир.
- Построение и обучение различных моделей машинного обучения для оценки стоимости недвижимости.
- Сравнение эффективности моделей и выбор наилучшего подхода для предсказания цен.

1 Анализ и обработка данных

Имеется датасет с объявлениями о продаже квартир в городе Новосибирске. В таблице всего 15000 строк, соответствующих разным квартирам, и 37 столбцов с признаками. Стоит отметить, что данные более качественные, чем в прошлой работе, это позволяет более явно указывать принцип отбора необходимых признаков и свойств для построения моделей. Для первичного обзора стоит рассмотреть все столбцы и количество уникальных значений для каждого из них (Табл. 1).

Таблица 1: Уникальные значения для каждого признака в таблице

Признак	**Количество**	**Признак**	**Количество**
Ссылка	15000	Долгота	3038
Точный адрес	3254	Широта	3111
Ценовая зона	10	Расстояние до центра	2924
Расстояние до метро	2921	Расстояние до остановки	2937
Расстояние до школы	2928	Расстояние до КАД	2
Заголовок	13318	Описание	12506
Тип рынка	2	Тип сделки	1
Город	1	Сегмент	1
Класс	2	Год постройки_	99
Год постройки	47	Высота поталков	1
Этажей	30	Этажей_	35
Количество комнат	7	Количество просмотров	70
Парковка	13	Отопление	7
Электричество	2	Водопровод	3
Канализация	4	Газ	4
Ремонт	4	Материал стен	8
Общая площадь, кв.м	2043	Жилая площадь, кв.м	616
Площадь кухни, кв.м	328	Цена, руб	2798
Удельная цена руб/кв.м	9710	Дата парсинга	112
Дата создания	91		

В сравнении с помещениями свободного назначения, у квартир больше полезных признаков, строк и лучше их заполненность, так же отсутствуют повторяющиеся значения квартир, поэтому производить поиск дубликатов не придётся. Сразу можно сказать, что для оценки не будут использоваться признаки: Ссылка, Точный адрес, Заголовок, Описание, Дата парсинга/создания, а также столбцы с единственным уникальным значением. Проведем анализ

категориальных признаков:

Таблица 2: Заполненность столбцов с категориальными признаками

Ценовая зона	**Количество**
Многоквартирная жилая застройка	7705
Исторический центр города	3365
Окраины	958
nan	787
Промзоны	713
ИЖС	687
Отсутствует	455
Зона автомагистралей	275
Зеленая зона	54
Центры деловой активности	1

Материал стен	**Количество**
nan	9850
Кирпичный	2244
Смешанный	967
Монолитный	635
Деревянный	492
Панельный	478
Блочный	330
Шлакоблоки	4

Как видно из таблицы, в данных превалирует ценовая зона многоквартирной жилой застройки, на неё приходится более половины строк, следом за ней идёт исторический центр города, а на остальные зоны приходится менее 20% квартир, но стоит выделить окраины и промзоны, так как далее будет видно, что для них средняя цена существенно ниже, что положительно скажется на точности оценки, если учесть это в модели. Заполнененность по признаку материала стен составляет меньше 30%, учёт этого признака не улучшает оценку моделей, следует не принимать его в расчёт.

Таблица 3: Заполненность столбцов с признаком "Парковка"

Парковка	**Количество**
nan	10391
Есть	2704
Подземная	1641
да	99
многоуровневая	69
предусмотрена	33
наземная	22
гостевая	17

Аналогично столбцу с материалом стен, наличие или тип парковки указаны менее чем для 30% данных, что не позволит построить точную оценку для этого признака, в модели этот столбец учитываться не будет.

Таблица 4: Заполненность столбцов с признаком "Тип рынка"

Тип рынка	**Количество**
Первичный	9189
Вторичный	5811

Тип рынка является очень важным признаком, так как он сильно коррелирует с ценой на квартиру. Для вторичного рынка средняя стоимость составляет 114 тысяч рублей, в то же время для первичного рынка цена приходится порядка 135 тысяч рублей.

Категориальные признаки, которые будут использоваться для дальнейшего анализа, представленны на гистограмме (Рис. 1).

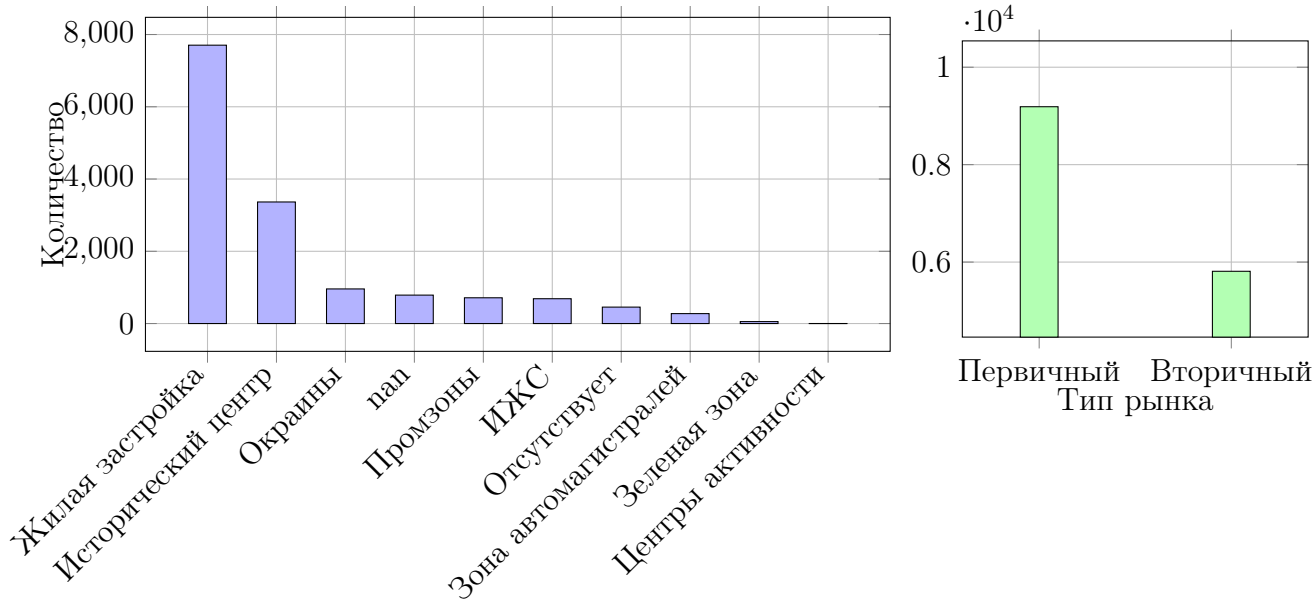

Рис. 1: Гистограммы распределения по категориальным признакам

Значения координат в столбцах 'Широта' и 'Долгота' можно преобразовать в более коррелирующий с ценой признак расстояния до центра города. Используя формулы гаверсинусов и заданные координаты центра города, для каждой строчки вычисляется новый признак. Для сравнения с предыдущей работой, средняя ошибка оценивания после этой замены уменьшилась на 1.5%, что является существенной величиной.

Для обеспечения устойчивости к выбросам для анализа и построения модели оценки следует оставить данные из интерквантильного размаха по самым важным численным признакам, как видно на (Рис. 2). После обработки осталось 13836 строк. В процентном соотношении было удалено гораздо меньше выбросов, чем при аналогичной процедуре для помещений свободного назначения. Стоит заметить, что данные отбрасываются только для построения

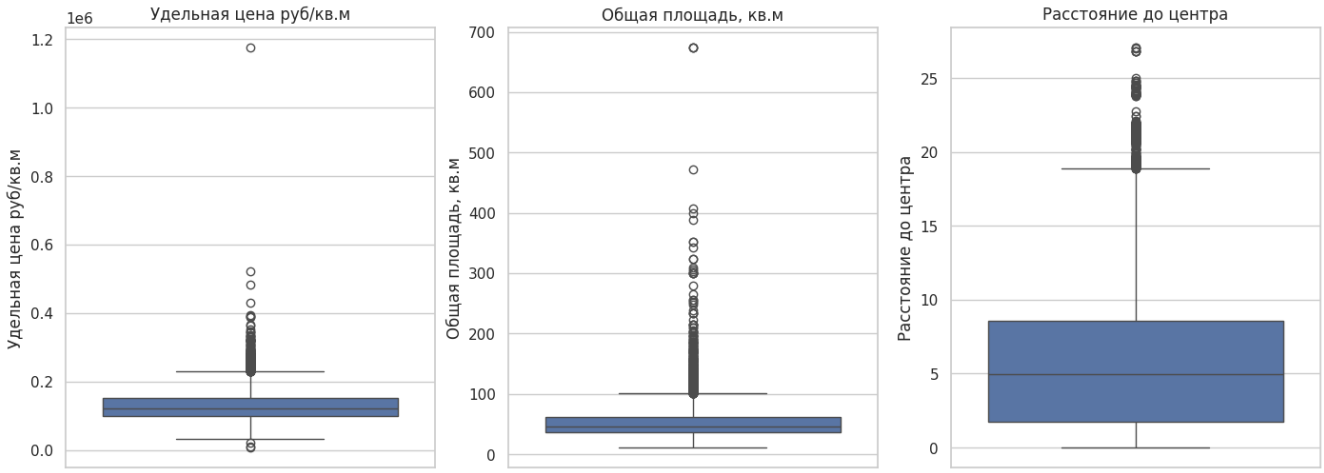

Рис. 2: Графики разброса выборки и интерквантильного размаха

оценки, но для проверки работы и качества оценивания можно использовать все данные, в том числе те, которые остались вне интерквантильного размаха.

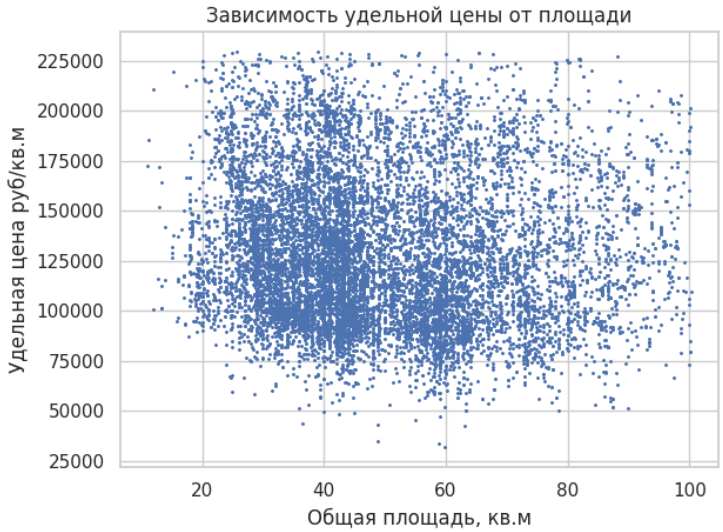

Рис. 3: График зависимости удельной цены от площади после удаления выбросов

На графике (Рис. 3) видно, что удельная цена распределена более равномерно, по крайней мере, в сравнении с таким же распределением в прошлой работе.

Построим гистограммы распределения числовых признаков после удаления выбросов.

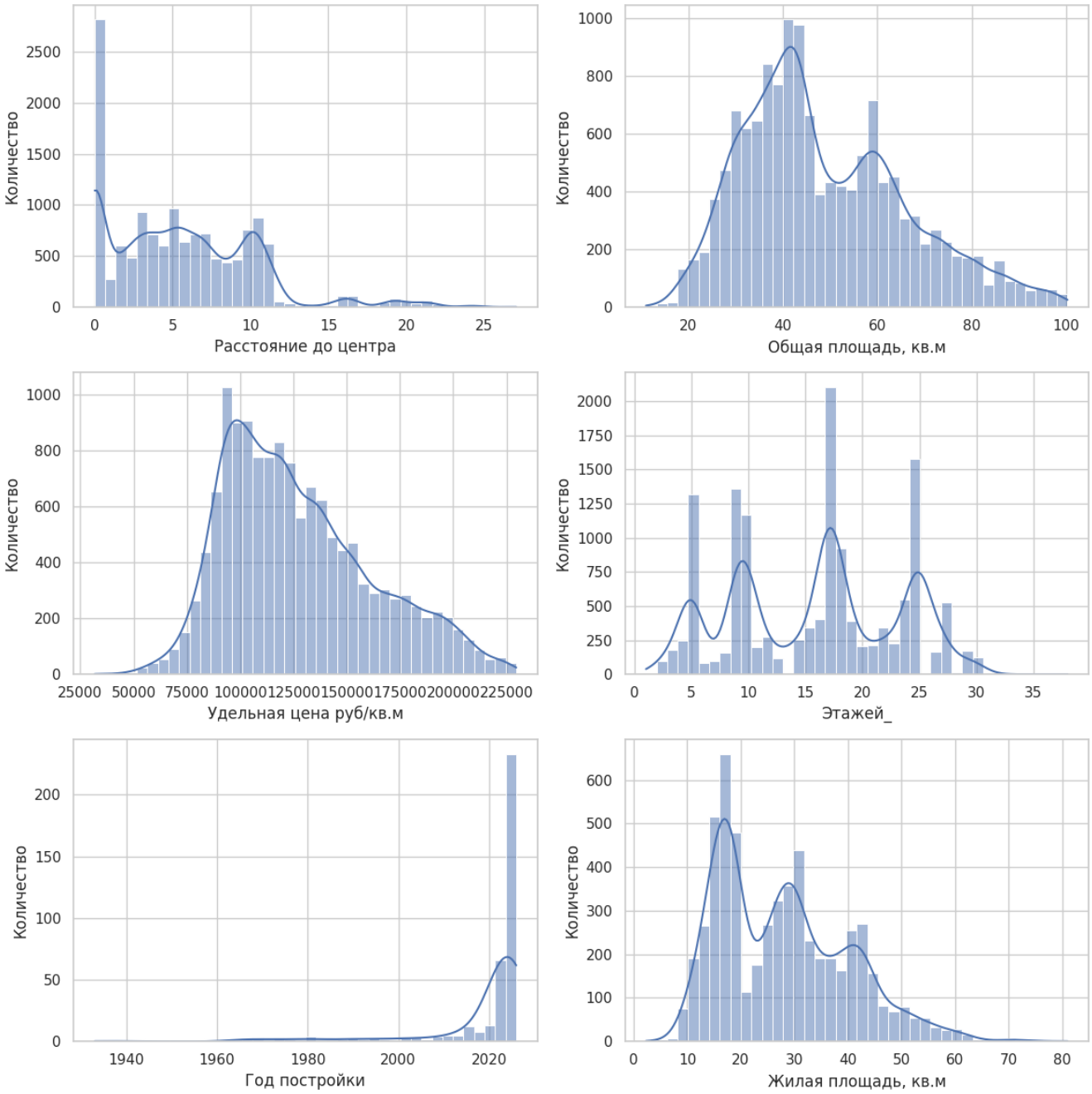

Рис. 4: Гистограммы распределения числовых признаков

Гистограммы численных признаков (Рис. 4) представляют собой большой интерес, по ним будет производиться переход к категориальным признакам. Каждый признак можно разбить на интервалы, которые станут категориями. Например, на графике общей площади заметно, что квартир площадью 30, 40 и 60 квадратных метров существенно больше, чем остальных. Эти площади коррелируют с количеством комнат в квартире, и в совокупности дают качественный категориальный признак. Аналогичный смысл несёт в себе распределение жилой площади, но она указана для малого количества квартир, поэтому при учитывании

её как признака в модели, улучшения оценки не было заметно. На гистограмме этажности выделяются пятиэтажные и девятиэтажные дома, как правило, они соответствуют рынку вторичной недвижимости и их цена ниже, в то время как дома более 10 этажей дороже, и чаще относятся к рынку новостроек. Год постройки указан для очень малого количества квартир, для оценки этот признак использоваться не будет.

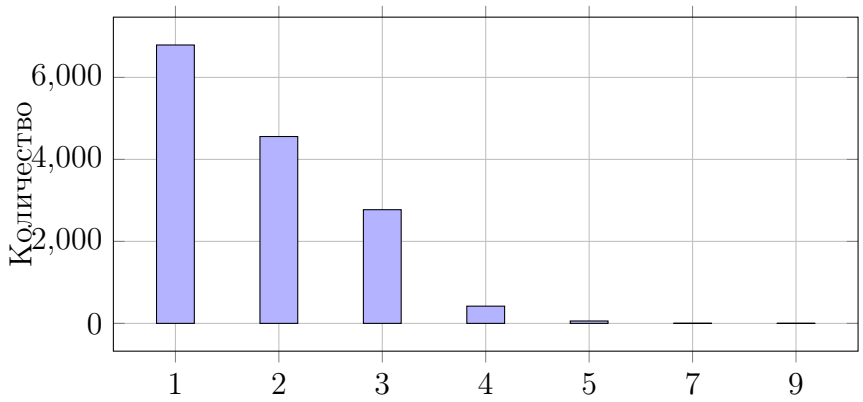

Рис. 5: Гистограмма количества комнат

Количество комнат тоже существенно влияет на цену, можно разбить на четыре категории: однокомнатные, двухкомнатные, трехкомнатные и четырех- и более комнатные квартиры.

Посмотрим как зависит цена от ценовой зоны:

Таблица 5: Статистические характеристики ценовых зон

Ценовая зона	**Среднее**	**Медиана**	**Максимум**	**Минимум**
Отсутствует	98343	95115	137500	81694
nan	113534	98647	228800	51839
Зеленая зона	141220	137959	205238	90000
Зона автомагистралей	174230	178183	220621	79187
ИЖС	137566	131000	224938	51388
Исторический центр города	133389	127272	229571	42944
Многоквартирная жилая застройка	125312	119109	229487	31772
Окраины	102727	102446	183333	48888
Промзоны	145928	148133	225993	62973

В таблице (Табл. 5) видно, что самой дорогой зоной является исторический центр города, после него идёт многоквартирная жилая застройка, и самой дешёвой является зона окраин. Остальные ценовые зоны представлены в очень малом количестве, поэтому их стоит объединить в единый категориальный признак.

2 Оценка стоимости недвижимости

Как и в случае с помещениями свободного назначения, сначала будет произведена оценка удельной стоимости квартир с помощью модели линейной регрессии, так же с регуляризацией Лассо и Ридж. Вдобавок к ним будут использованы методы k-ближайших соседей и дерево решений.

Модель k-ближайших соседей оценивает стоимость на основе k ближайших объектов в обучающей выборке. Для вычисления расстояния между объектами используется евклидово расстояние:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n} (x_j - x_{ij})^2} \tag{1}$$

Находим k объектов в обучающей выборке, расстояние до которых минимально. Стоимость объекта x оценивается как среднее значение целевой переменной для k ближайших соседей:

$$\hat{y}(x) = \frac{1}{k} \sum_{i=1}^{k} y_i \tag{2}$$

Дерево решений разделяет пространство признаков на области, используя правила принятия решений. На каждом узле дерева выбирается признак и пороговое значение, которые максимизируют критерий информации. Как критерий информации для регрессии используется снижение дисперсии:

$$\Delta \mathrm{D}(S, a) = \mathrm{D}(S) - \left(\frac{|S_L|}{|S|} \mathrm{D}(S_L) + \frac{|S_R|}{|S|} \mathrm{D}(S_R) \right), \tag{3}$$

где S — множество объектов в узле, S_L и S_R — множества объектов в левом и правом подмножествах после разделения по признаку a, D - дисперсия.

Процесс повторяется рекурсивно для каждого дочернего узла до выполнения критерия остановки. Оценка стоимости в листе дерева производится как среднее значение целевой переменной объектов, попавших в этот лист:

$$\hat{y}(x) = \frac{1}{|S_{\text{leaf}}|} \sum_{i \in S_{\text{leaf}}} y_i \tag{4}$$

Результат работы модели:

```
MdAPE для линейной регрессии: 15.28%
MdAPE для Лассо регрессии: 15.29%
MdAPE для Ридж регрессии: 15.29%
Среднеквадратическое отклонение: 867427656.23
R-квадрат: 0.30
```

Линейная регрессия показала себя на 5.5% лучше чем на выборке с помещениями свободного назначения, в том числе из-за тех оптимизаций, которые были применены в данной работе. Разные регуляризации не улучшили результат.

K-NN:

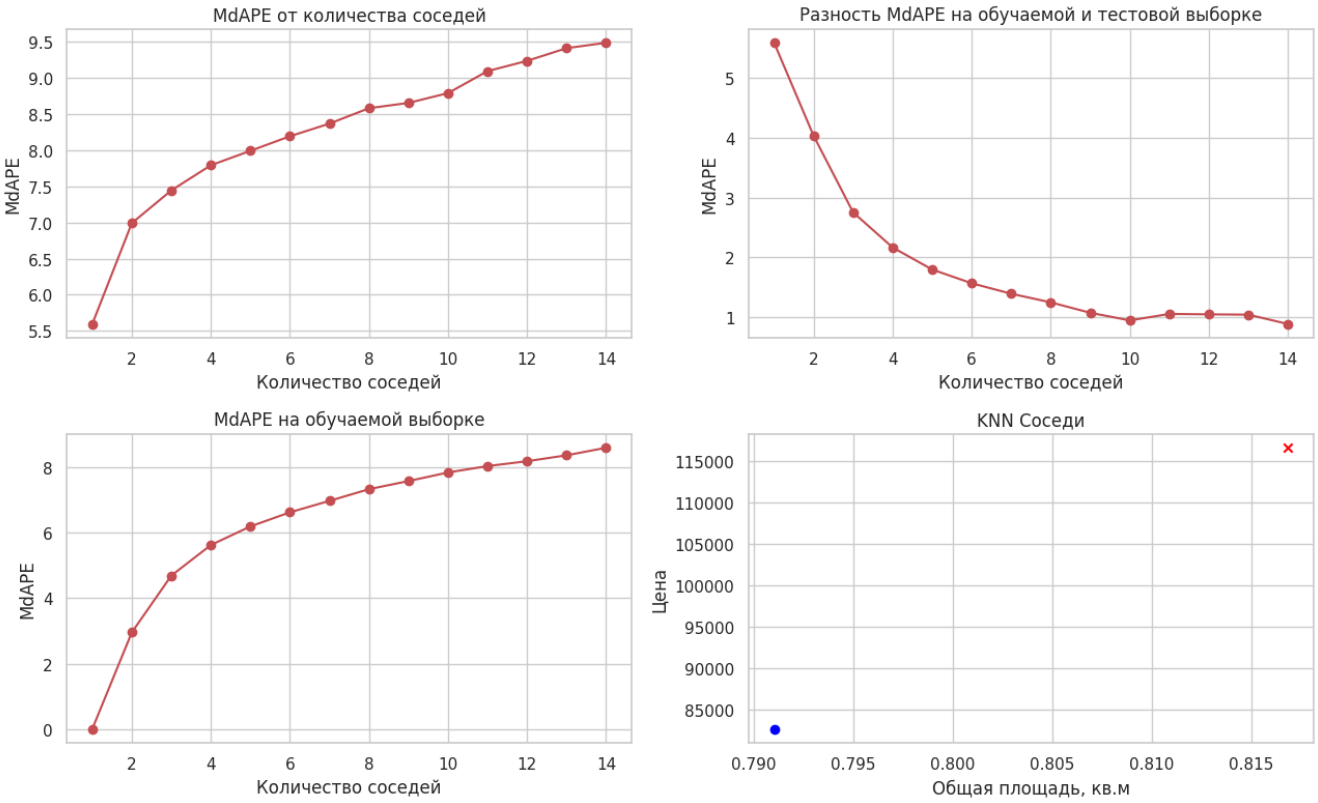

Рис. 6: Графики зависимости средней ошибки для модели k-ближайших соседей

На Рис. 6 отображена зависимость MdAPE от количества соседей. Для обучаемой выборки она начинается в 0 для 1 соседа. Это связано с тем, что для любого элемента ближайшим соседом всегда является он сам же, что в случае наличия 1 соседа даёт идеальное предсказание. Значение ошибки на тестовой выборке имеет минимум в 5.5%, но такая модель просто даёт цену известного и ближайшего по расстоянию в пространстве параметров элемента обучаемой выборки. Такая модель будет плохо себя вести для выборсов, квартир в новых ценовых сегментах и тому подобных случаях, когда ближайший сосед в обучаемой выборке может оказаться слишком далеко. Более надёжной моделью будет такая модель, где разность ошибки на тестовой и обучаемой выборке перестанет убывать. В нашем случае это 10 соседей. Эта модель имеет следующие результаты:

```
Среднеквадратическое отклонение: 508400682.25
R-квадрат: 0.59
MdAPE на обучающих данных: 7.53%
```

```
MdAPE на тестовых данных: 8.77%
```

Этот результат в 2 раза лучше чем любая линейная модель, которая была использована ранее. Но этот результат можно улучшить ещё сильнее, если построить метод, в котором сам элемент тестовой выборки не будет учитываться при обучении. Это позволит брать за оценку не среднее значение k соседей, а брать значения с весами, зависящими от расстояния до соседа.

Дерево принятия решений:

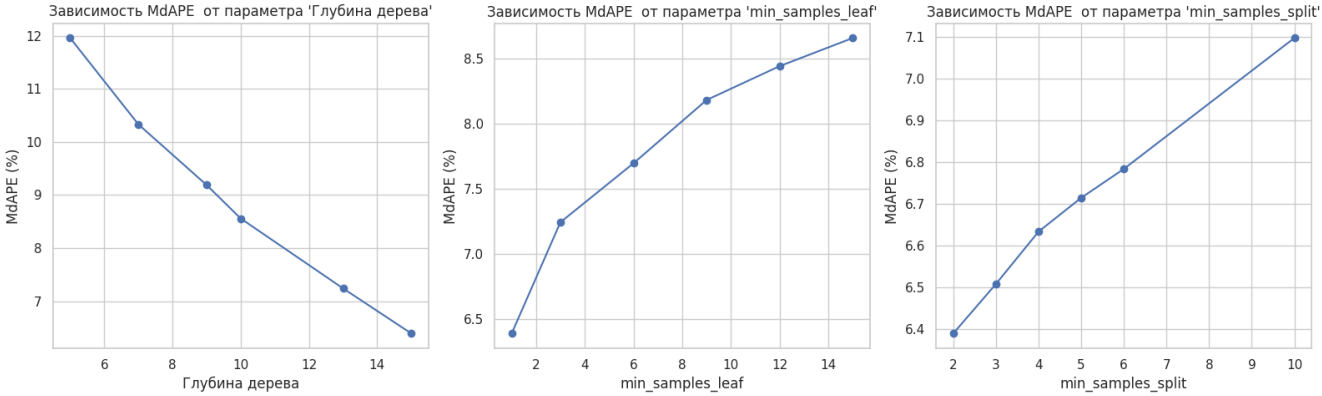

Рис. 7: Графики зависимости средней ошибки для дерева принятия решений

У дерева принятия решений есть 3 параметра: глубина рекурсии, количество элементов выборки чтобы создать лист дерева, количество элементов для разделения листа.

Чем больше глубина и меньше листья, тем больше возможных подмножеств выборки можно сгруппировать, поэтому графики (Рис. 7) имеют такой вид. Стоит ограничить глубину дерева до 10, так как преодолев это значение происходит переобучение модели, аналогично тому, что было описано для модели k-ближайших соседей. Дерево принятия решений показывает следующие результаты:

```
Среднеквадратическое отклонение: 442229835.49
R-квадрат: 0.64
MdAPE на обучающих данных: 7.01%
MdAPE для дерева принятия решений: 8.20%
```

Лучшей моделью оказалось дерево принятия решений.

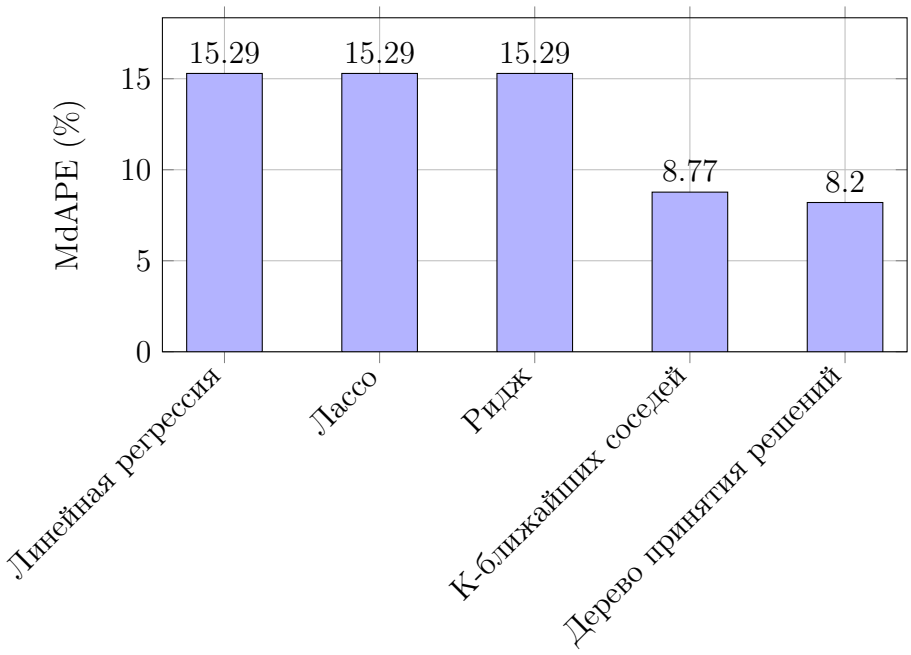

Рис. 8: Сравнение моделей по MdAPE

Заключение

Результатом этой работы является новый взгляд на оценку стоимости квартир с использованием математической статистики. Можно сделать вывод, что при оценке большую роль играют сами данные, их распределение, качество признаков. В прошлой работе были рассмотрены только линейные модели, из-за чего оценки не учитывали нелинейные зависимости в стоимости. Рассмотрев метод K-ближайших соседей и дерево принятия решений удалось устранить этот недостаток, что улучшило оценку в два раза. Полученная оценка имеет среднюю ошибку 8.2%, что является очень хорошим результатом.

Список литературы

1. Колмогоров, А.Н. Теория вероятностей и математическая статистика / А.Н. Колмогоров. - М.: Наука, 2005. - 581 с.

2. Крамер, Г. Математические методы статистики / Г. Крамер. - М.: Мир, 1975. - 648 с.

3. Кремер, Н. Ш. Теория вероятностей и математическая статистика : учебник и практикум для вузов / Н.Ш. Кремер. — М.: Юрайт, 2023. — 538 с.

4. Федоткин, М.А. Лекции по анализу случайных явлений / М.А. Федоткин. - М.: Физматлит, 2016. – 464 с.

5. Хасти, Т., Тибшериани, Р., Фридман, Д. Основы статистического обучения. Интеллектуальный анализ данных, логический вывод и прогнозирование / Т. Хасти, Р. Тибшериани, Д. Фридман. – М.: Вильямс, 2020. – 768 с.